Learning Indoors Free-space Segmentation for a Mobile Robot from Positive Instances

Christos Sevastopoulos[1,2], Joey Hussain[1], Qiyuan An[1],
Stasinos Konstantopoulos[2], Vangelis Karkaletsis[2], Fillia Makedon[1]
[1]Department of Computer Science & Computer Engineering, University of Texas at Arlington, Arlington, TX, USA
[2] Institute of Informatics and Telecommunications, NCSR "Demokritos", Agia Paraskevi, Attica, Greece

{christos.sevastopoulos,joey.hussain,qxa5560}@mavs.uta.edu
{konstant,vangelis}@iit.demokritos.gr
makedon@uta.edu

2023 Seventh IEEE International Conference on Robotic Computing (IRC) | 979-8-3503-9574-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IRC59093.2023.00010

***Abstract*—Accurate indoor free-space segmentation is a demanding task due to the intricate and dynamic nature of indoor environments. We propose an indoors free-space segmentation method that associates large depth values with navigable regions. Our method employs an unsupervised masking technique that, using positive instances, generates segmentation labels based on textural homogeneity and depth uniformity. Using the estimated free-space masks and a Dense Prediction Transformer (DPT) feature representation, a SegFormer model is fine-tuned on our custom-collected indoor dataset. Our experiments demonstrate sufficient performance in complex scenes where the identification of free space is challenging.**

***Index Terms*—Computer Vision, Free-space segmentation, Deep Learning, Mobile Robots**

I. Introduction

Robust free-space segmentation is a fundamental task for mobile robots operating indoors. By correctly identifying areas devoid of obstacles, robots can navigate safely and efficiently, while avoiding collisions. However, accomplishing accurate free-space segmentation remains challenging due to the complex and dynamic nature of indoor environments. Challenges such as the stochastic nature of the environment, presence of humans, variations in lighting conditions and occlusions, can largely affect the perception of free space. Additionally, indoor scenes with ambiguous and semantically diverse objects, like furniture, household items, and interior structures may exhibit textures similar to traversable surfaces, and thus lead to inconsistent predictions.

In this article, our objective is to overcome these limitations by proposing a novel learning-based approach for indoor free-space segmentation. Our method goes beyond conventional object-centric models [1], [2] by leveraging depth information to focus on precise detection and segmentation of free space. By utilizing the power of pre-trained transformer-based networks, transfer learning, and positive instances, our aim is to train a model that can identify safe and navigable areas within complex indoor environments. In order to achieve this, we propose that areas classified as traversable free-space are likely to exhibit greater depth values.

As an overview, the contributions of this work are the following:

- We hypothesize that areas classified as traversable free-space exhibit greater depth values
- An automated masking technique that leveraging textural homogeneity and depth uniformity
- Learning to predict indoor free-space segmentation using only positive instances for training

II. Related Work

Extensive research has been conducted on predicting a scene's geometry using RGB-D information, primarily focusing on outdoor scenarios [3], providing valuable insights into identifying the location of traversable free-space [4]. In indoor settings, notable attempts have been made to determine the scene's traversability only at a higher conceptual level [5] without specifically delving into the segmentation of free-space. The integration of RGB and depth information has demonstrated its advantages in the field of semantic segmentation. Cao et al. [6] exploited estimated depth features to develop an RGB-D semantic segmentation approach that incorporates a multi-task training strategy. Nekrasov et al. [7] propose a method that adapts a real-time semantic segmentation network to handle multiple tasks with asymmetric datasets.

With the emergence of transformers, the *Vision Transformer (ViT)* architecture [8] has been associated with an array of computer vision tasks due to its excellence at capturing global context and long-range dependencies within the input image. SegFormer [9] combines the ViT-architecture with lightweight *Multilayer Perceptron (MLP)* decoders, and has shown significant efficiency for semantic segmentation tasks. Other transformer-based models highlighting the diverse applicability of such in computer vision tasks can be found in [10] with particular focus on image restoration and in [11] addressing image classification, object detection, and semantic segmentation.

III. Methodology

Figure 1 illustrates the proposed methodology's architecture. Our approach is based on the central hypothesis that areas classified as traversable free-space will generally exhibit

979-8-3503-9574-7/23/$31.00 ©2023 IEEE
DOI 10.1109/IRC59093.2023.00010

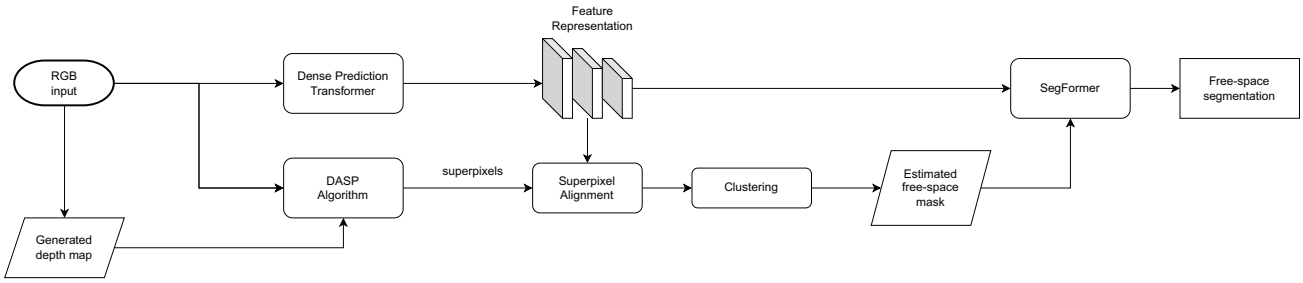


Fig. 1: Proposed methodology


Fig. 2: Illustrative examples of positive and challenging instances. For the positive example (left) the robot is free to traverse while in the challenging case (right), the free space is limited due to the presence of cluttered objects

greater depth values. To validate this hypothesis, we focus on training a refined semantic segmentation model using positive RGB indoor instances that the robot can safely traverse, such as hallways and wide paths. For generating the depth maps, we utilized the depth estimation model provided by Niantic Labs [12], allowing to extract valuable depth information from the aforementioned RGB instances. Additionally, we propose an unsupervised mask generation framework that leverages depth information to capture textural homogeneity and depth uniformity as indicators of free space. Subsequently, we evaluate the performance of the model on challenging scenarios where free space is present yet restricted due to the presence of cluttered objects (Figure 2).

To generate annotation masks for free-space segmentation, we build upon the following main hypothesis: Free space is expected to be depicted by textural homogeneity in conjunction with depth uniformity. This means, that in order to search for free space in the image, we consider these particular superpixels that are characterized by similarity in both color and depth information. Therefore, we use a superpixel oversegmentation algorithm to generate unsupervised free-space masks by capturing intra-region similarities towards producing meaningful segments: the Depth-Adaptive Superpixels (DASP) algorithm [13]. Overall, our method is influenced by the findings and approaches outlined in [14], however we differentiate in the sense that ”seed” superpixels are generated through the guidance of depth information. In lieu of empirically computing the parameters that define the location prior, we use the DASP algorithm to generate the seeds needed for segmentation.

We intend to associate superpixel clusters with features that have been extracted by a Dense Prediction Transformer [15], that uses the *Vision Transformer* (ViT) architecture as a backbone for dense prediction tasks such as semantic segmentation. We remove the last layer of the pre-trained Intel's DPT-Large[1] model, and the resulting extracted features, in the form of a vector of size 577x1024, exhibit rich high-level semantic information as well as spatial details at a higher resolution.

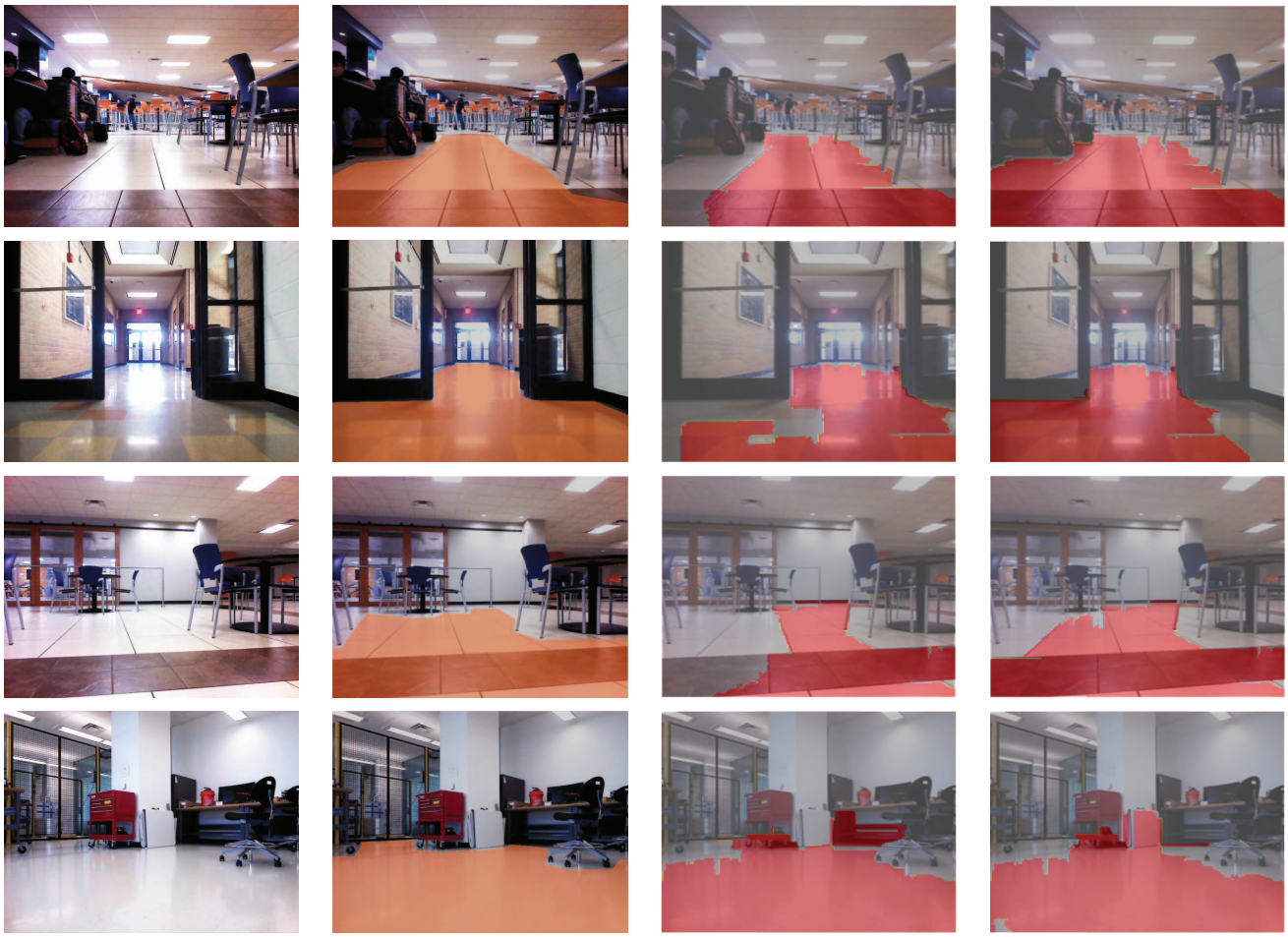
To establish a spatial association between the distinct features derived from the DPT and the coherent superpixels generated by the DASP algorithm, we use a sophisticated technique known as *superpixel alignment*, inspired by the work by Tsutsui et al. [14] and influenced by the concept of RoIAlign [16]. This alignment is crucial in ensuring that the information extracted by the DPT accurately corresponds to the meaningful regions represented by the superpixels. Afterwards, we employ average pooling to aggregate the features within each superpixel and therefore generate a comprehensive representation of the superpixels' characteristics.

We aim to guide the segmentation towards traversable areas that exhibit the largest depth. Therefore, we cluster superpixels with respect to the semantic features and additionally, we select the cluster that corresponds to the area with the largest depth. As highlighted by Weikersdorfer et al. [13], the density of superpixel clusters in the image space is computed from the depth image. For detailed information on how this density is computed, the reader is referred to the original article. Consequently, our objective is to identify the clusters within the Poisson disk distribution that are most likely to define free space. For a given depth image, we estimate the superpixel density values, and using the k-means clustering technique, we single out the clusters. Ultimately, we use a SegFormer-B0 pre-trained on ImageNet-1K, and replace the classification head for fine-tuning on our dataset.

IV. Experimental Setup

The Summit-XL Steel robot was deployed in various buildings on the University of Texas, Arlington (UTA) campus, where it navigated hallways and areas with changing lighting

[1]https://github.com/isl-org/DPT



(a) Input Images (b) Ground truth (c) Prediction with RGB , [14] (d) Prediction with RGB+Depth

Fig. 3: Illustrative examples of the method's performance, last row depicts erroneous prediction

conditions and diverse objects. In total, we collected a dataset of 3324 monocular RGB images. Our primary objective was to utilize the Positive-Unlabeled (PU) method [17], to divide each source set (corresponding to a specific university building) into two subsets: positive and unlabeled image instances. The annotation process is the one described in [18]. From the unlabeled set, we manually removed scenes where the robot had limited or no free space to proceed, and labeled scenes with both free space and obstacles as "challenging". Thus, our dataset for experiments consisted of 2553 images, including 2010 positive and 443 challenging instances .

V. Results and Discussion

A. Implementation Details

We used the PyTorch[2] framework as the basis of our experiments. Training was done on a machine with 2 Titan RTX GPUs (24GB GDDR6 RAM, 4608 Cuda Cores). Using the standard cross-entropy loss function, we trained for 50 epochs unless an early stopping callback terminated the trial upon observed convergence. As training parameters we used: batch size = 16, learning rate = 0.01 and weight decay = 5e-4. The fine-tuning of the layers was accomplished using stochastic gradient descent (SGD). All images' initial dimensions of 640x480 pixels were re-scaled to 224x224 using the default PyTorch interpolation. In the following sections, we present the best results achieved when using 1800 of positive instances for training, and 453 challenging instances for testing.

[2]https://pytorch.org

B. Performance Analysis

We train our method on the positive instances and test on the challenging ones. For our experiments, the standard semantic segmentation metrics, Intersection over Union (IoU) is used and we found that the best results were achieved for k=5 clusters. Table I provides a summary of the performance of different methods under various input and model configurations in terms of IoU scores. Overall, the results show that incorporating depth information improves the accuracy of free-space segmentation compared to using only RGB images. We

observe that our method, using the aforementioned unsupervised annotation framework, outperforms the corresponding SegNet-based method from [14] while also maintaining comparable performance versus its supervised counterparts.

C. Qualitative results

We compare the performance of our method, that uses both RGB and depth information during training to generate masks, against the method presented by [14] and only uses RGB (Figure 3). Overall, it is noted that including depth information enhances the accuracy of free-space prediction compared to relying solely on RGB data. That is because depth information can provide valuable cues for free-space segmentation in the narrow passages along with some depth discontinuities, that help distinguishing the objects' edges and boundaries from the surrounding free-space regions.

SegFormer's proficiency to effectively capture the long-range dependencies and spatial information, along with the features of the pretrained DPT within the input data, is of paramount importance. It enables the proposed algorithm to differentiate between free-space regions and obstacles, even in challenging scenarios with objects, doors, and varying textures. Nonetheless, with respect to the fourth row of Figure 3, it is noted that the tile positioned against the wall leads to a misunderstanding of texture, because it resembles a traversable surface that the algorithm has been previously trained on.

TABLE I: Performance of each method given different inputs

Method	RGB	Depth	Training Data	Model	IoU Score
[14]	✓	×	Automated	SegNet	0.773
[14]	✓	✓	Automated	SegNet	0.801
Ours	✓	×	Automated	SegFormer	0.813
Ours	✓	✓	Automated	SegFormer	0.844
Ours	✓	✓	Hand-labeled	SegFormer	0.878
Ours	✓	✓	Automated	SegNet	0.824
Ours	✓	✓	Hand-labeled	SegNet	0.862

VI. Conclusions and Future Work

This paper explores the significance of linking navigable regions with areas of higher depth for the purpose of free-space segmentation. We use an efficient automated masking technique, which utilizes textural homogeneity, depth uniformity and positive scenes to generate meaningful segments. Moreover, we fine-tune a SegFormer on our custom-collected dataset, training on positive and testing on challenging instances, witnessing satisfactory performance. Avenues of future research include explore a more explicit combination between textural and geometric features as well as experimentation with a larger and more diverse dataset.

Acknowledgment

The authors would like to thank Vasileios Kitsakis for his valuable insight and feedback.

References

[1] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 525–11 538, 2020.

[2] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, "Conditional object-centric learning from video," *arXiv preprint arXiv:2111.12594*, 2021.

[3] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European Conf on computer vision (ECCV)*, 2018, pp. 36–53.

[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conf on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[5] N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese, "GONet: A semi-supervised deep learning approach for traversability estimation," in *Proceedings of 2018 IEEE/RSJ Int'l Conf on Intelligent Robots and Systems (IROS)*, 2018.

[6] Y. Cao, C. Shen, and H. T. Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 836–846, 2016.

[7] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, "Real-time joint semantic segmentation and depth estimation using asymmetric annotations," in *2019 Int'l Conf on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7101–7107.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[10] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conf on computer vision and pattern recognition*, 2022, pp. 17 683–17 693.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF Int'l Conf on computer vision*, 2021, pp. 10 012–10 022.

[12] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Int'l Conf on computer vision*, 2019, pp. 3828–3838.

[13] D. Weikersdorfer, A. Schick, and D. Cremers, "Depth-adaptive supervoxels for rgb-d video segmentation," in *2013 IEEE Int'l Conf on Image Processing*. IEEE, 2013, pp. 2708–2712.

[14] S. Tsutsui, T. Kerola, S. Saito, and D. J. Crandall, "Minimizing supervision for free-space segmentation," in *Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 988–997.

[15] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF Int'l Conf on computer vision*, 2021, pp. 12 179–12 188.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE Int'l Conf on computer vision*, 2017, pp. 2961–2969.

[17] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," *Machine Learning*, vol. 109, pp. 719–760, 2020.

[18] C. Sevastopoulos, M. Theofanidis, A. Hebri, S. Konstantopoulos, V. Karkaletsis, and F. Makedon, "Indoors traversability estimation with rgb-laser fusion," in *2023 IEEE 19th Int'l Conf on Automation Science and Engineering (CASE)*. IEEE, 2023, pp. 1–7.